Structured speaker variability in Japanese stops: relationships within versus across cues to stop voicing

James Tanner,[1] Morgan Sonderegger,[1] and Jane Stuart-Smith[2]

[1)] *Department of Linguistics, McGill University, Montreal, QC, Canada*

[2] *University of Glasgow, Glasgow, UK*[a]

A number of recent studies have observed that phonetic variability is constrained across speakers, where speakers exhibit limited variation in the signalling of phonological contrasts in spite of overall differences between speakers. This previous work has focused predominantly on controlled laboratory speech and on contrasts in English and German, leaving unclear how such speaker variability is structured in spontaneous speech and in phonological contrasts which make substantial use of more than one acoustic cue. This study attempts to both address these empirical gaps and expand the empirical scope of research investigating structured variability by examining how speakers vary in the use of positive voice onset time and voicing during closure in marking the stop voicing contrast in Japanese spontaneous speech. Strong covarying relationships within each cue across speakers are observed, whilst between-cue relationships across speakers are much weaker, suggesting that structured variability is constrained by the language-specific phonetic implementation of linguistic contrasts.

[a)] james.tanner@mail.mcgill.ca;

I. INTRODUCTION

The acoustic realisation of segments varies substantially across languages, phonological contexts, and speakers. Within a single language, the realisation of a particular segment can differ as a function of phonological context (Cho and Ladefoged, 1999), speech rate (Allen et al., 2003), and many other linguistic and social factors (e.g., Foulkes et al., 2001). Individual speakers may differ in the realisation of speech sounds because of numerous factors: some speakers are more prone to hyperarticulation of segments (Johnson et al., 1993; Lindblom, 1990), differ in their anatomical characteristics (Peterson and Barney, 1952), or simply arrive at different acoustic targets as function of probabilistic approximation of the speech sounds in their community (Bybee, 2001; Pierrehumbert, 2001). This kind of speaker-level variability poses a potential challenge for the perception of speech (Kleinschmidt, 2018), where the mapping from values in a multi-dimensional acoustic space to abstract phonological categories (e.g., [+voice], [-high], etc.) is differently realised for individual speakers (Liberman et al., 1967; Lisker, 1986). How, then, do speakers successfully convey the presence of singular linguistic categories despite individual variation in those categories' realisations? One way in which individual variability may be constrained is by the existence of underlying *structure* in the realisation of speech sounds across speakers: namely, that speakers' individual productions are related in a way that is fundamentally non-random. For example, whilst speakers vary in the realisation of a single acoustic parameter such as Voice Onset Time (VOT) for stops, the differences between individual speakers' VOT values for different places of articulation are highly correlated (Chodroff and Wilson, 2017; Hullebus et al.,

2018). Speakers may also show similar kinds of structured variation across *multiple* cues to the production of a speech sound, evidenced by observed covariation in VOT and F0 across voiced and voiceless stops (Bang, 2017; Chodroff and Wilson, 2018; Clayards, 2018; Schultz et al., 2012).

Beyond one study on Scottish English (Sonderegger et al., 2020) and two studies on American English (Chodroff and Wilson, 2017, 2018), most recent research on structured variation across individuals has focused on production in controlled laboratory speech, either isolated words or reading sentences (Chodroff and Wilson, 2017; Clayards, 2018; Hullebus et al., 2018; Schultz et al., 2012). The phonetic realisation of stop contrasts is known to be 'enhanced' in laboratory speech relative to conversational speech (Baran et al., 1977; Lisker and Abramson, 1967) – for example voiced/voiceless VOT differences are larger – and so it is less clear how variability is structured in less-controlled speech. Examining spontaneous speech alongside more controlled speech may provide new insights into structured speaker variability in phonetic realisation, as for other aspects of speech, such as variability in vowel production (DiCanio et al., 2015; Gahl et al., 2012; Meunier and Espresser, 2011). Our understanding of structured speaker variability is also largely derived from research which has examined languages such as English and German, which primarily use VOT to signal a range of contrasts in word-initial stops (e.g., Lisker and Abramson, 1964, 1967). How speakers vary in languages where the stop contrasts involve the use of additional phonetic cues is not well-understood.

This study addresses these theoretical gaps by focusing on the acoustic realisation of stops in spontaneous Japanese. Japanese uses both *positive* VOT - the period encompassing the

duration of aspiration and the stop burst - and the presence of voicing in the stop closure for marking the contrast between voiced and voiceless stops (Shimizu, 1996; Tsujimura, 2014, Section II A). Typically 'VOT', in work on Japanese and other languages, is defined as the time between the release of the stop and onset of glottal pulsing for the following vowel: VOT is positive if voicing begins after the release of the stop closure, and negative otherwise. In that definition, VOT is both an indirect measure of 'burst duration' and aspiration (when positive) and the presence of voicing during the closure (when negative). In this study, which focuses on structured variability, it is important for us to capture the complex interplay between laryngeal and supralaryngeal actions/timing in Japanese stops through two dimensions. In line with several recent studies which distinguish between positive VOT and the presence of voicing during closure (Kim et al., 2018; Kleber, 2018; Seyfarth and Garellek, 2018; Sonderegger et al., 2020), we use the term 'pVOT' to refer to the duration of 'burst plus aspiration' following the release of the closure. We use 'voicing during closure' (VDC) to refer to any voicing throughout the stop closure. The Japanese stop voicing contrast has been observed to be changing through the decreased use of voicing during closure, resulting in a system more like an English-style aspiration contrast (Takada, 2011; Takada et al., 2015), and so may provide insight into how speakers vary in the use of both pVOT and voicing during stop closure, as well as in how both parameters are used to realise the voicing contrast. This study expands the search for structured speaker variability by examining the evidence for three kinds of such structure across speakers of spontaneous Japanese: (1) *within* a phonetic cue across different voicing categories (e.g., pVOT between voiced and voiceless stops); (2) the size of the voicing contrast *across* cues across categories

(i.e., the relative difference between voiced and voiceless stops); and (3) *across* phonetic cues within voicing categories (i.e., the relationship between pVOT and voicing during closure in voiced and voiceless stops).

II. BACKGROUND

A. Acoustic cues to stops & stop voicing

VOT as traditionally defined, is well-established as the primary acoustic cue for the stop voicing contrast in a range of languages where voiced stops have shorter average VOT than their voiceless counterparts (Abramson and Whalen, 2017; Liberman et al., 1958; Lisker and Abramson, 1964). Japanese maintains a two-way stop voicing contrast, distinguishing between 'voiced' {/b/, /d/, /g/} and 'voiceless' {/p/, /t/, /k/} categories: acoustically, Japanese voiced stops may be realised either with prevoicing (negative) or short-lag VOT (Gao and Arai, 2019; Nasukawa, 2005; Shimizu, 1996), and voiceless stops are realised with a VOT intermediate between short ('unaspirated', Tsujimura, 2014) and long-lag ('moderately aspirated', Riney et al., 2007; Shimizu, 1996). Whilst less is known about variability in Japanese stops, much work has focused on how stops are modulated in English: here it is assumed that these factors are to some extent language-independent and are thus also relevant for Japanese stops. Stop VOTs are affected by a range of linguistic factors, such as place of articulation (Docherty, 1992; Lisker and Abramson, 1964), preceding phoneme manner (Docherty, 1992; Yao, 2009), vowel height (Klatt, 1975), phrasal position (Cho and Ladefoged, 1999; Kim et al., 2018; Lisker and Abramson, 1964; Yao, 2009), and speech rate

(Allen et al., 2003). Most work on English VOT has used controlled speech, though the few studies which have looked at English spontaneous speech have confirmed a robust difference in VOT between voiced and voiceless stops (Baran et al., 1977; Sonderegger et al., 2017; Stuart-Smith et al., 2015). These studies focused on variation between groups of speakers; few studies have examined individual speaker variation in spontaneous English stops (Chodroff and Wilson, 2018; Sonderegger et al., 2020), leaving unaddressed questions concerning variability between individual speakers in languages with different phonetic implementations for stops (Section II B).

The degree of vocal fold vibration during the closure (Lisker, 1986), reflected in our VDC measure, is much less studied than VOT, though English voiced stops are more likely to contain VDC than their voiceless counterparts (Docherty, 1992; Sonderegger et al., 2020). Most research on VDC has focused on English read speech (e.g., Davidson, 2016, 2018; Kim et al., 2018). For both voiced and voiceless stops, VDC is more likely in phrase- or word-medial contexts (Docherty, 1992; Lisker and Abramson, 1964, 1967). VDC in phrase-initial stops, sometimes referred to as 'negative VOT', has been observed for English (Hunnicutt and Morris, 2016; Lisker and Abramson, 1964, 1967) and other languages (Abramson and Whalen, 2017). Additionally, VDC is more likely when the preceding segment is voiced (Davidson, 2016, 2018; Docherty, 1992), also in spontaneous speech (Sonderegger et al., 2020). With the exception of geminated consonants, all syllables in Japanese are either open (ending in a vowel) or have a nasal coda (Tsujimura, 2014): all segments preceding stops in these cases are underlyingly voiced, then, and this should affect the likelihood of a stop being realised with VDC. Closure voicing is also used as a contrastive cue for voicing

in Japanese, though recent studies have shown that the prevoiced variant of the voiced stop has become less common in phrase-initial position (Gao and Arai, 2019), and may represent a sound change towards the exclusive use of positive VOT coupled with F0 variation to signal the voicing contrast (Gao and Arai, 2019; Gao et al., 2019; Kong et al., 2014; Takada, 2011).

B. Individual speaker variability in stops

Differences between individual speakers have been noted since the earliest acoustic studies of stop production (e.g., Lisker and Abramson, 1964). As opposed to being random variation, these differences between speakers are highly structured: speaker differences in VOT are consistent after controlling for other linguistic factors, such as speech rate (Allen et al., 2003; Theodore et al., 2009). Speaker mean VOTs for different places of articulation in voiceless stops have been shown to be highly correlated in both English (Chodroff and Wilson, 2017) and German (Hullebus et al., 2018): despite overall differences in a given speaker's mean VOT, realisation of the contrasts between voiceless stops (i.e., /p/ ∼ /t/, /p/ ∼ /k/, /t/ ∼ /k/) exhibits strong linear relationships. With respect to speaker variability across *multiple cues* to stop production, Chodroff and Wilson (2018) show that American English speakers covary in use of three cues (VOT, F0, and spectral centre of gravity), and Glaswegian English speakers covary in the relationship between positive VOT and the degree of VDC (Sonderegger et al., 2020). Similar relationships exist between VOT and F0 in marking the laryngeal contrast in English, German, and Korean (Bang, 2017; Schultz et al., 2012), whilst Schertz et al. (2015) observed speaker differences in the correlated use of VOT, F0,

and closure duration in L2 English-Korean speakers, and Clayards (2018) reported similar findings for VOT, F0, and following vowel duration in English.

In order to characterise the sources of structured variability within an individual's phonological grammar, Chodroff and Wilson (2017, 2018) propose a 'principle of uniformity'. Uniformity in this sense seems to refer to a linear relationship in the acoustic production of two segments across speakers; the degree of variation in the difference between two speech sounds across speakers is constrained such that the realisation of one sound has a predictive relationship with the other. Whilst speakers may vary in their overall use of a given phonetic cue (i.e., where that speaker is situated on this line), the relative difference between two segments with respect to that parameter is consistent across speakers. Much of the evidence for Chodroff & Wilson's proposition of uniformity is derived from studies of English, which uses an aspiration-based phonetic implementation of stops.

By examining the structure of speaker variability in spontaneous Japanese, a new language with a different phonetic implementation of voicing, we can consider further possible evidence for phonetic uniformity in a new empirical setting. This examination takes two forms here: the first considers how speakers modulate the stop voicing contrast within a given phonetic cue (pVOT or Voicing During Closure). The second concerns how these two cues are manipulated together in signalling this contrast. Whilst some research has examined speaker variability across multiple cues, especially in English (e.g., Chodroff and Wilson, 2018; Clayards, 2018), the predictions are less clear for a language like Japanese where the cues to stop voicing differ from English and where a number of possibilities exist. For example, if pVOT and Voicing During Closure share an intrinsic articulatory link, we could

expect strong correlations between pVOT and Voicing During Closure, such that speakers with more aspirated stops also produce less Voicing During Closure. This would correspond to the intuition behind the traditional 'VOT' measure, that stop production is often well-characterized by a single dimension (Abramson and Whalen, 2017, a closure voicing–degree of aspiration continuum). Alternatively, the lack of an intrinsic link between the cues may result in no observed correlations between the respective use of pVOT and Voicing During Closure. These questions also address how phonetic uniformity across speakers might be constrained and whether such constraints may relate to language-specific properties.

III. METHODS

A. Data

The data used here comes from the Core subset of the Corpus of Spontaneous Japanese (CSJ, Maekawa et al., 2000), constituting approximately 45 hours of speech recorded 1999-2001 from 137 speakers (58 female), born between 1930 and 1979. Within the CSJ, speaker birth years are grouped into increments of 5 years (e.g., 1930-34, 1935-39, 1940-44, etc); in order to ensure sufficient numbers of speakers per group, speakers were allocated into groups of 10 years (1930-39, 1940-49, etc). The variety of Japanese in the CSJ is 'Common' Japanese: a standard variety that derives many of its linguistic features from the Tokyo dialect (Maekawa et al., 2000). Each recording is approximately 30 minutes long, and is predominantly academic interviews and informal public speaking, though a subset (approximately 5%) is conversational dialogue and reading passages. The Core subset contains

extensive phonetic and prosodic annotation, including hand-corrected segmental boundaries, presence of vowel devoicing, and voice quality (Kikuchi and Maekawa, 2003). Relevant for the measures taken here, stops were annotated for (1) onset of stop closure, (2) stop burst – the first transient spike – and (3) the onset of the vowel. The segmentation criteria for the hand correction are provided in Fujimoto et al. (2006): for our purposes, onset of following vowel was determined by CSJ annotators as the beginning of periodicity for the vowel (Fujimoto et al., 2006, p.330); see Figure 1. The annotations also noted whether the stop was fully realised, defined by whether a clear closure, burst, and voice onset could be visually observed (the CSJ does not contain annotation for negative VOT).

In order to ensure that stops examined in this study were fully realised, certain stops were excluded from further analysis: any stop marked as not having a clear closure and burst (56,661 tokens); stops followed by a devoiced vowel, as voicing onset could not be ascertained (11,939 tokens); stops immediately following hesitations (11,991 tokens); geminate stops (19,785 tokens), as geminates in Japanese are not phonologically contrastive for voicing in native words and often devoice (Kawahara, 2015); stops from word-medial contexts (72,681 tokens), as stops reduce in these contexts (Cho and Ladefoged, 1999; Kim et al., 2018); and stops from non-spontaneous read speech (4,790 tokens). Prosodic position is defined in the corpus using the X-JToBI prosodic-labelling scheme (Maekawa et al., 2002), which numerically represents the perceived strength of a prosodic juncture through 'Break Indices' (BIs). BI labelling is based on a range of perceptual cues including segmental lengthening, F0 reset, and changes in voice quality (Venditti, 2005). Junctures with a BI value of 1 typically represents a word boundary within an Accentual Phrase (AP), BI value of 2 represents

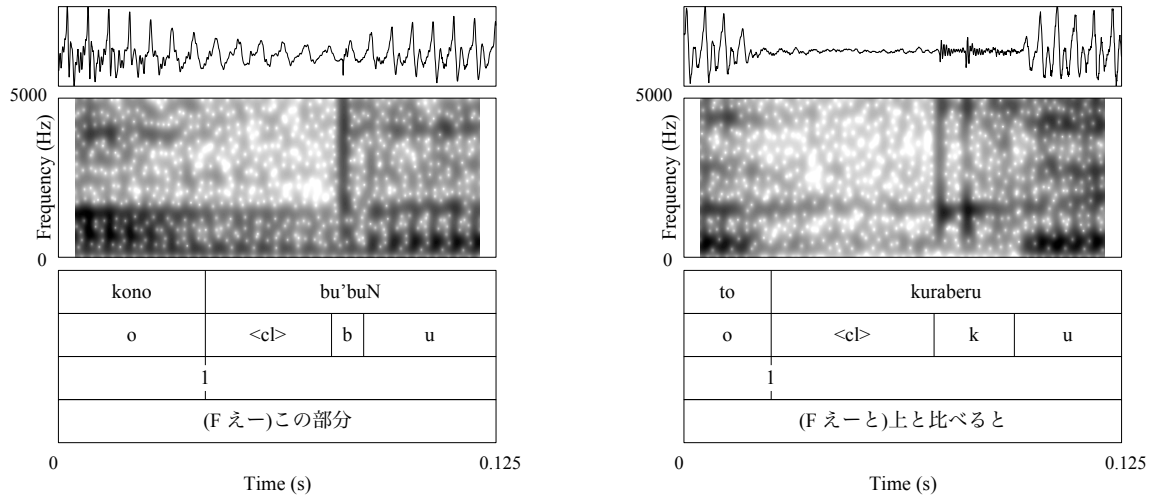


FIG. 1. Waveforms and accompanying annotations for phrase-internal stops realised with and without voicing during closure ('kon**o** **b**ubun', (a); '**to** **k**uraberu', (b), respectively) produced by a female speaker taken from a 125ms time window. Closure annotated as <cl>. Top tier represents word-level transcription, second tier contains phone & sub-phone annotations, third tier marks prosodic boundaries via Break Index, and fourth tier contains utterance transcription.

the boundary between two APs, whilst BI values of 3 indicate the edge of an Intonational Phrase (IP). We excluded all tokens with *no* BI value (which are predominantly word-medial). The final set of stops analysed therefore constitutes word-initial stops excluding potentially-problematic cases.

B. Voicing during closure (VDC)

The goal of the VDC measure is to characterise the presence of voicing during closure, which plays a key part in signalling phonological voicing in Japanese. It is well known, however, that realisation of voicing within the stop closure is more complicated in connected speech than that in isolated words (Abramson and Whalen, 2017; Lisker and Abramson,

1964, 1967). Voicing may continue for the entire stop closure ('full voicing'), or may subside ('bleed') and/or return just prior to the release ('trough') (Davidson, 2016). Cases like this make the traditional definition of 'negative VOT' difficult for characterising the voicing pattern. Davidson (2016, 2018) observed that voicing during closure corresponding to negative VOT in American English appeared in only a handful of tokens. Whilst several studies have focused on negative VOT in laboratory speech (Gao and Arai, 2019; Gao et al., 2019; Kong et al., 2014; Takada, 2011; Takada et al., 2015), no work to our knowledge has examined stop closure voicing patterns in Japanese connected speech similar to Davidson (2016, 2018) for English.

Davidson (2016) notes the likelihood of voicing during closure in English is closely tied to the voicing of the preceding segment: preceding voiced segments (vowels, sonorants) are more likely to induce voicing during closure than voiceless segments. This is important here since *all* preceding segments are voiced: Japanese syllables are either open (i.e., consonant-vowel) or contain a nasal coda (Tsujimura, 2014): as geminated stops are excluded, all stops are preceded by a vowel or a nasal (potentially with an intervening pause). A preceding vowel does not guarantee the realisation of voicing in the stop closure, however: Figure 1 (a) shows a voiced stop with voicing throughout the stop closure ('full voicing'), whilst no such voicing during closure is evident in a voiceless stop in the same phonetic context (Figure 1, (b)).

Our goal for the VDC measurement is to characterise the presence of phonetic voicing during closure in terms of the likely presence of an active voicing gesture (Beckman et al., 2013). In order to capture this, the presence of VDC is defined in binary terms between the

presence or *absence* of active voicing during closure. This aims to exclude common cases of passive voicing which are often short (less than 20ms) and weak in amplitude, in contrast to an active voicing gesture, characterised by clear periodic voicing for a substantial portion of the closure and the presence of pitch. This deviates from previous studies on English using similar approaches (Davidson, 2016; Sonderegger et al., 2020) where voicing during closure was trichotomised into 'no', 'partial', or 'full' voicing, determined by the relative portion of the observed voicing within the closure. The decision to use a binary voicing distinction in this study was based on the goal of restricting to cases where an active voicing target was clearly present or not, as well as on the empirical observation that both Davidson (2016) and Sonderegger et al. (2020) found that effects were more apparent in their respective binary ('no' versus 'full') models than comparing relative degrees of voicing. Our characterisation of VDC as distinct from pVOT enables both voicing presence and pVOT to be examined as independent cues to stop production: given observations that it is possible for speakers to produce stops with both voicing during closure and pVOT (Abramson and Whalen, 2017; Kim et al., 2018; Sonderegger et al., 2020), it is important to know if speakers are able to modulate both pVOT and voicing during closure independently to signal the Japanese stop voicing contrast.

In order to calculate a measure of VDC, both the mean F0 and the 'fraction of unvoiced frames' were extracted from the labelled stop closure using Praat Voice Report (Boersma and Weenink, 2017). As Voice Report has been known to produce inaccurate measurements of voiced frames when viewed using the Editor window, our calculations followed Eager (2015): specifically, the Voice Report was produced by a Praat script without using the Editor

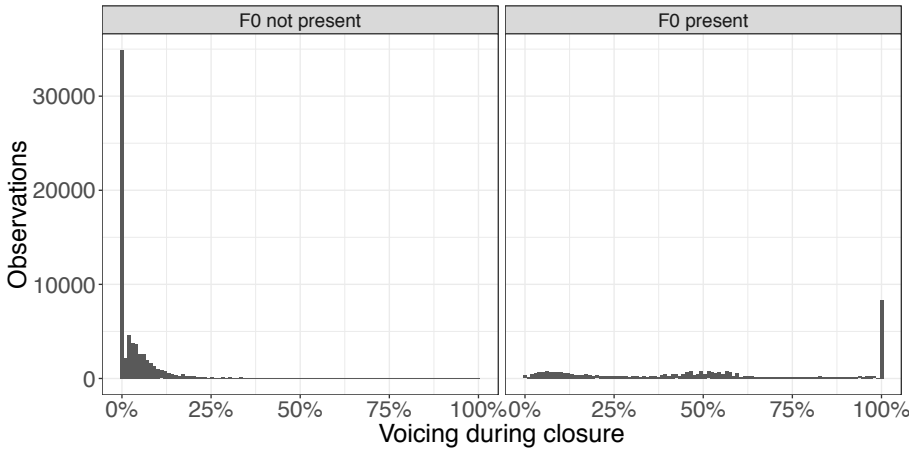


FIG. 2. Histograms showing the distribution of the percentage of voicing during closure by whether F0 was also detected within the stop closure. 100 bins used within each histogram, meaning that each bar represents 1%.

window, using gender-specific pitch ranges (70-250Hz for males; 100-300Hz for females), and a time step of 0.001 seconds. The percentage of voicing during closure was calculated by subtracting 100 from Voice Report's proportion of the interval with *no* voicing: for example, if Voice Report returned an unvoiced closure value of 66%, then voicing % $= 100 - 66 = 34$.

Our main goal involved determining which instances of stop voicing were most likely produced with an active voicing gesture. For the purposes of this study, tokens which satisfied two criteria were analysed. The first was whether F0 was present in the closure; the second was whether a significant portion of the closure contained voicing. Numerous values have been proposed in the literature for what proportion of the closure reflects active voicing, such as 'greater than 50%' (Abramson and Whalen, 2017) and 'greater than 10%' (Davidson, 2016). Here, decisions regarding the cutoffs were determined by examining the distribution of voicing during closure percentages with and without the presence of F0. As

shown in Figure 2, voicing during closure with no accompanying F0 (left panel) ranges from 0% to approximately 15%, and so VDC (reflecting an active voicing gesture) was considered to be absent for such tokens. When F0 is present (right panel), a large number of tokens exhibited 100% voicing during closure with a small cluster around 50%. To include these tokens, the 'present' VDC category was defined as tokens with the presence of F0 and at least 35% voicing in the closure. Other cases were taken to indicate that voicing was unreliable: F0 may have been present but the lack of substantial voicing % suggests potential voicing bleed. Unreliable tokens were excluded (18,960; 17.5%), meaning that all remaining tokens are assumed to be realised with either no voicing during closure or an active voicing gesture. Our final dataset used for analysis contained 90,160 tokens (3,440 types) from 137 speakers (58 female), with an average of 658 tokens per speaker (range of tokens per speaker: 149–2,913).

C. Models

The goal of this study is to examine evidence for structured speaker variability (1) within individual acoustic cues; (2) in the voicing contrast across cues across voicing categories; and (3) across cues within individual phonetic categories. In order to address these questions, pVOT and VDC were statistically modelled to characterise individual speaker differences whilst controlling for a range of factors known to influence both cues (Section II A). pVOT (log-transformed)[1] and VDC were jointly modelled using a multivariate Bayesian mixed model using *brms* (Bürkner, 2018), an R front-end for the Stan programming language (Carpenter et al., 2017). A Bayesian model returns a *distribution* of potential values for all

model parameters, which makes it possible to estimate correlations across speakers as well as the uncertainty associated with each correlation. This is ideal for addressing all three research questions, as the strength of relationships across speakers can be characterised formally in terms of both the strength of the correlations and the range of possible correlations consistent with the data. As pVOT and VDC are fit within the same model, it is possible to also directly estimate the speaker correlations *across* phonetic cues, which is crucial for research questions (2) and (3). Finally, the use of a statistical model to estimate speaker correlations, rather than estimating correlations from empirical data as in most previous work on structured speaker variability, allows for correlations (and individual speaker values for each cue) to be estimated whilst controlling for the range of other factors known to affect both pVOT and VDC (Sec. II A).

The model consists of a sub-model predicting pVOT and a sub-model predicting VDC, and terms linking these sub-models together. We first describe the terms in each sub-model, which were identical. Each sub-model included the following population-level ('fixed-effect') predictors for stop **voicing**, previous phoneme **manner**, speaker **birth year** and **gender**, stop **place of articulation**, speech **style**, prosodic **position**, log-transformed word **frequency**, speaker **mean** and **local** (relative to mean) speech **rate** (Sonderegger et al., 2017; Stuart-Smith et al., 2015), the presence of a preceding **pause**, and following vowel **height**. To control how each predictor influenced the realisation of the voicing contrast, two-way interation terms between stop voicing and all other predictors were also included in the model. Continuous predictors (speaking rates, frequency, vowel duration) were centred and divided by two standard deviations (Gelman and Hill, 2007). Two-level factors (voicing,

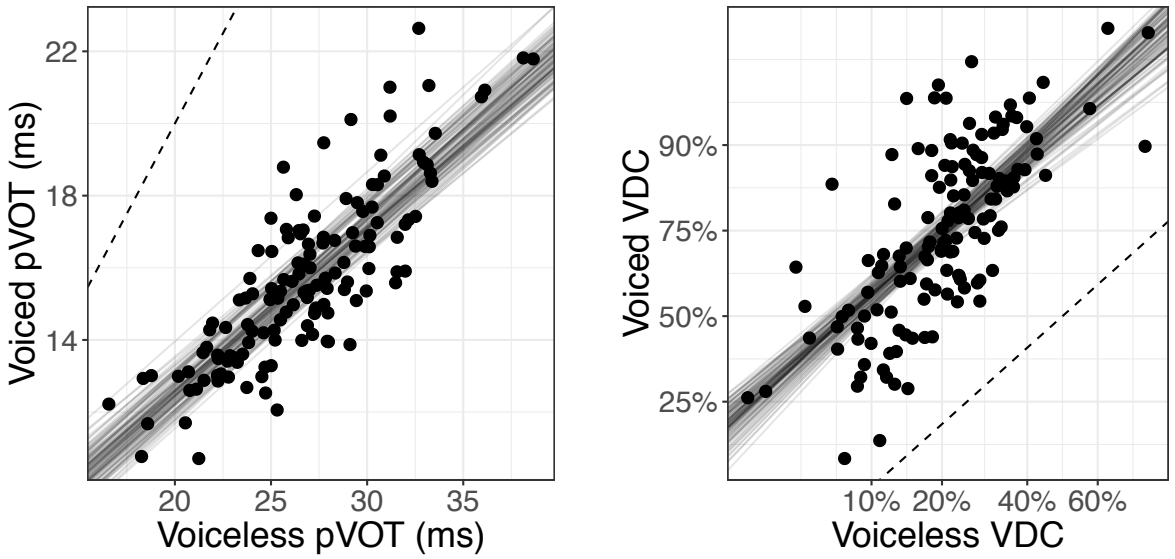
accent, gender, vowel height, pause) were converted into binary (0/1) measures and centred. Predictors with three or more levels (birth year, place of articulation, phoneme manner) were coded with sum contrasts. For group-level ('random-effect') predictors, the model was fit with a random intercept for words; speaker-level effects consisted of a random intercept and random slopes for all population-level predictors (with the exception of style, age, and gender). As the relationship between a speaker's overall value for pVOT/VDC and the size of their voicing contrast is of direct interest, both models included a correlation term between the speaker-level intercept and the voicing predictor. The pVOT and VDC sub-models were tied together by three correlations between the key speaker-level effects: intercepts, voicing, and the correlation between them. For example, the correlation term between the pVOT intercept and the VDC intercept captures the extent to which speakers with higher mean pVOT are more likely to use VDC. The model used 8000 samples across 4 Markov chains and was fit with weakly-informative 'regularising' priors (Nicenboim and Vasishth, 2016; Vasishth et al., 2018) of normal distributions with a mean of 0 and standard deviations of 1 and 0.5, and 0.5 for pVOT intercept, VDC intercept, and fixed effect parameters respectively. The default prior in *brms* for group-level effects was used: a half Student's t-distribution with 3 degrees of freedom and a scale parameter of 10. Correlations used the LKJ prior (Lewandowski et al., 2009) with $\zeta = 2$, in order to give lower prior probability to perfect (1/-1) correlations, as recommended by Vasishth et al. (2018).[2] All data and code used is available at `https://osf.io/grw25/`.

IV. RESULTS

The research questions concern the relationships observed across speakers *within* each cue (1) as well as *across* both cues (2, 3), and so correlations were calculated for each of the 8000 draws from the posterior sample and reported as the median, 95% credible interval (CrI), and the posterior probability of the parameter not including 0, using `fitted_draws` and `median_qi`, respectively, from the *tidybayes* package (Kay, 2019). Speaker-level variability is first examined *within* pVOT and VDC separately (IV A) before examining the relationships *between* both cues across speakers (IV B). Following Nicenboim and Vasishth (2016), we consider there to be strong evidence for a non-null effect if the 95% CrI for the parameter does not include 0; if 0 is within the 95% CrI but the probability of the parameter not changing direction is at least 95%, this is considered to represent weak evidence for a given effect. Crucially the strength of evidence for an effect is distinct from its magnitude, and so the strength of a given predictor's effect on pVOT/VDC is considered alongside its relative evidence. The size or magnitude of a given correlation is assessed in terms of Cohen's conventions (Cohen, 1988): correlations with sizes between 0 and 0.1 (in either direction) are considered to be *negligible*; those with sizes between 0.1 and 0.3 to be *small*; between 0.3 and 0.5 to be *medium*; and *strong* correlations have values larger than 0.5. Cohen's conventions are heuristic and should be considered relative to previous effect sizes observed for a given phenomenon. Given the relative scarcity of results on the relationships across speakers, Cohen's conventions provide some initial benchmarks against which to evaluate the relative relationships within and across phonetic cues.

Correlation	ρ	95% CrI	$\Pr(\rho <> 0)$
Voiceless pVOT, Voiced pVOT	0.77	[0.709, 0.821]	1
Voiceless VDC, Voiced VDC	0.664	[0.594, 0.729]	1

TABLE I. Median correlation, 95% credible intervals (CrI), and posterior probability of within-cue correlations (Spearman's ρ) across speakers sampled from the model posterior with all other predictors held at their 'average values' (e.g., mean word frequency, mean across all places of articulation, etc).



(a) (b)

FIG. 3. Model-estimated cue values for pVOT (a) and VDC (b) for voiceless (x-axis) and voiced (y-axis) stops. One point is the posterior mean value for a particular speaker. Black lines are 100 lines of best fit drawn from the model posterior to show direction and uncertainty in the correlation. Dashed line is $y = x$, where the value for voiceless stops equals that for voiced stops. pVOT plot in linear (millisecond) scale; VDC plot is in logit-scaled probability scale to illustrate differences at extreme upper and lower probabilities.

A. Within-cue variability

The effects of the population-level parameters on pVOT were as expected, including the size of the voicing contrast (Table III in Appendix). As the pVOT voicing contrast is maintained across all population-level effects (i.e., no parameter neutralised or reversed the basic voiceless > voiced pattern, including speaker age) and speaker-level variability is of primary interest for our research questions, these parameters provide controls for the speaker-level variability; the fixed effects are not discussed further. Figure 3 (a) demonstrates the strong correlation between speakers' voiced and voiceless pVOTs (95% CrI = [0.709, 0.821]; Table I, row 1): each point represents a speaker's median estimated voiceless (x-axis) and voiced (y-axis) pVOT value. All individual speakers have higher pVOTs for voiceless than voiced stops, indicated by all points appearing on one side of the dashed $y = x$ line. Speakers differ in their particular pVOT values, but the relative difference between their voiced and voiceless pVOTs (i.e., the voicing contrast) is consistent: the regression lines demonstrate this linear relationship, where speakers both maintain the contrast between stops, and speakers with long pVOTs for voiceless stops also have long pVOTs for voiced stops.

No population-level effect neutralised or reversed the VDC voicing contrast (Table IV, in Appendix), meaning that VDC is always predicted to be more likely for voiced than voiceless stops ($\hat{\beta} = 2.99$, CrI = [2.76, 3.21], $\Pr(\hat{\beta} > 0) = 1$). Note, however, the large effect of the presence of a preceding pause on VDC, which suggests that speakers producing spontaneous Japanese are substantially less likely to produce VDC directly following a pause ($\hat{\beta} = -3.24$,

	Correlation	ρ	95% CrI	$\Pr(\rho < > 0)$
Voicing contrast	pVOT contrast, VDC contrast	0.198	[−0.001, 0.346]	0.974
Within-category	Voiced pVOT, Voiced VDC	−0.348	[−0.423, −0.27]	1
	Voiceless pVOT, Voiceless VDC	0.135	[0.038, 0.228]	1
Across-category	Voiceless pVOT, Voiced VDC	−0.152	[−0.233, −0.066]	0.99
	Voiced pVOT, Voiceless VDC	0	[−0.092, 0.093]	0.5

TABLE II. Median correlation, 95% credible intervals (CrI), and posterior probability of across-cue correlations (Spearman's ρ) across speakers sampled from the model posterior with all other predictors held at their 'average values' (e.g., mean word frequency, mean across all places of articulation, etc). pVOT contrast = voiceless pVOT − voiced pVOT; VDC contrast = voiced VDC − voiceless VDC.

CrI = [−3.51, −2.97], $\Pr(\hat{\beta} < 0) = 1$), consistent with experimental findings (Gao and Arai, 2019). Comparing across voicing categories, Figure 3 (b) shows that speakers maintain a strong positive relationship between their voiced and voiceless VDCs (95% CrI = [0.594, 0.729]; Table I, row 2). No speaker has a reversed voicing contrast for VDC, reflected by all speaker values (represented as points) appearing above the $y = x$ line. The consistent positive slope of the regression lines illustrate that, as with pVOT, speakers who are more likely to produce VDC for voiced stops are also more likely, on average, to produce voiceless stops with VDC.

B. Across-cue variability

Having shown above how speakers vary *within* a single cue (pVOT, VDC) between voiced and voiceless stops (question 1) we now address whether speakers vary *across* cues in pro-

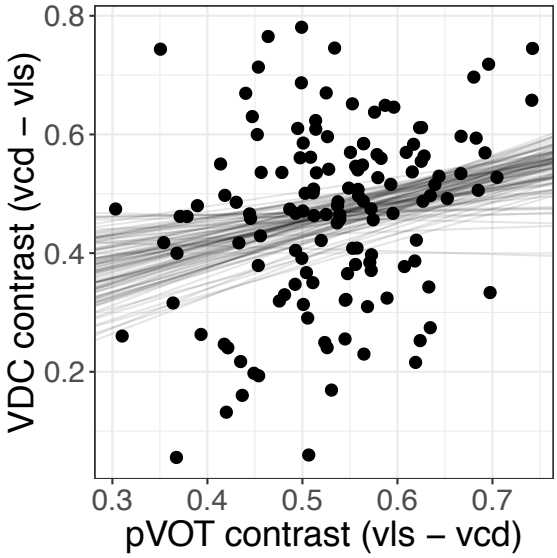


FIG. 4. Model-estimated voicing contrast sizes for pVOT (x-axis) and VDC (y-axis). Each point is the posterior mean for a particular speaker. Black lines are 100 lines of best fit drawn from the model posterior to show direction and uncertainty in the correlation.

duction, where speakers may coordinate both cues in signalling the stop voicing contrast (question 2), or within a voicing category (question 3). Comparing the size of the voicing contrast for each cue, a weak positive relationship across speakers can be observed (95% CrI = [−0.001, 0.346]; Table II, row 1): this can be interpreted as meaning that the voicing contrast sizes across cues are somewhat linked, with speakers differing in precisely how they realise the voicing contrast simultaneously across both pVOT and VDC (Figure 4).

Given the strong correlations across speakers in single use of a given cue (Figure 3) and the observation that speakers only weakly vary in the size of their voicing contrast across both cues (Figure 4), the question remains as to how speakers covary in the use of pVOT and VDC within specific phonetic categories. In other words, do speakers' values for one cue (e.g., pVOT) within a category (e.g., voiceless stops) correlate with their values for the

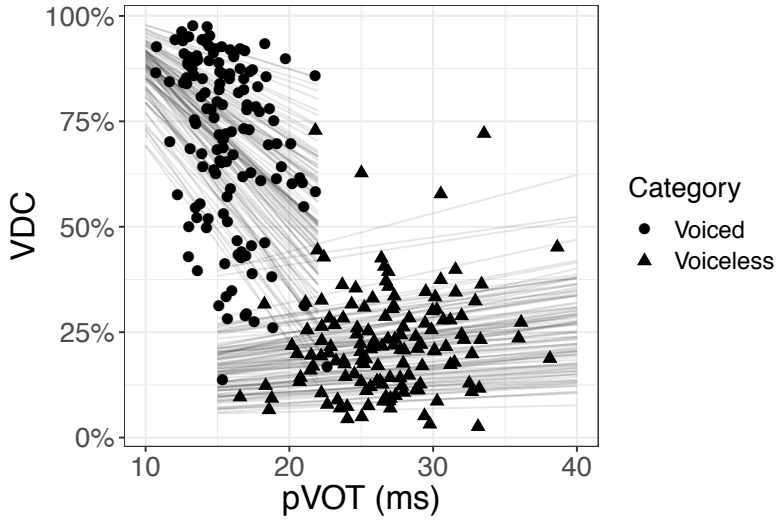


FIG. 5. Model-estimated cue values for pVOT (x-axis) and VDC (y-axis). Voicing category of the stop is represented by shape (points = voiced; triangles = voiceless). Points and lines represent the same values as in Figures 3 and 4.

other cue (VDC) in that same category? Figure 5 demonstrates this combination of cues by voicing categories, and illustrates an asymmetry in the pVOT-VDC relationship between voiced and voiceless stops. Speakers provide strong evidence for a negative relationship of medium strength between pVOT and VDC in voiced stops (Figure 6, a), meaning that speakers with larger voiced pVOTs have a lower voiced VDC likelihood (95% CrI = [−0.423, −0.27]; Table II, row 2). For voiceless stops, however, there is strong evidence for a weak *positive* relationship (95% CrI = [0.038, 0.228]; Figure 6, c; Table II, row 3). A negative relationship is also observed between speakers' voiced VDC rate and their voiceless pVOTs, though this is much smaller in magnitude than the voiced pVOT-voiced VDC relationship (95% CrI = [−0.233, −0.066]; Figure 6, d; Table II, row 4); voiceless VDC does not show

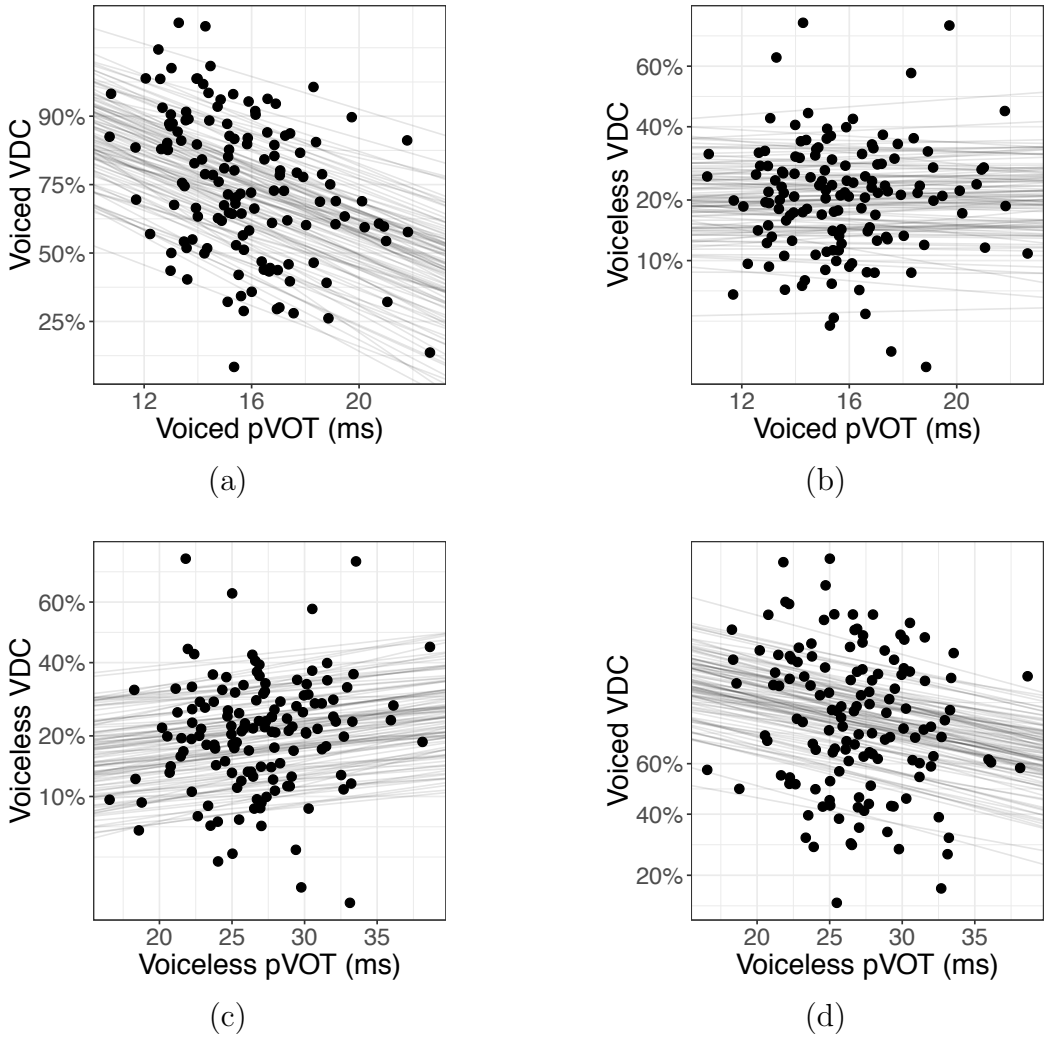


FIG. 6. Model-estimated cue values for pVOT (x-axis) and VDC (y-axis), comparing relationship between cues either within (left) or across (right) a given stop category. Points and lines represent the same values as in Figures 3 and 4. pVOT in linear (ms) scale; VDC in logit-scaled probabilities to show differences at extreme probabilities (near 0% or 100%).

409 a meaningful correlation with voiced pVOT across speakers (95% CrI = [−0.092, 0.093];

410 Figure 6, b; Table II, row 5).

V. DISCUSSION

The phonetic realisation of segments differs across languages, dialects, phonetic contexts, and individual speakers. Recent research has observed that this variability across individual speakers is *structured*: whilst speakers may differ in the overall value of a particular phonetic cue, they may demonstrate covariation in the use of one or more cues to mark linguistic contrasts (e.g., Chodroff and Wilson, 2018; Sonderegger et al., 2020; Theodore et al., 2009). Little is known about how speaker variability may be structured in a languages which show different phonetic and phonological signalling of linguistic contrasts. This study begins to address these empirical gaps by examining positive VOT and VDC as cues to stop voicing in spontaneous Japanese. Strong within-cue relationships are observed across speakers between voiced and voiced stops: whilst speakers differ in their overall values of pVOT or VDC, speakers are consistent in the relative *difference* within pVOT or VDC in marking the voicing contrast. These within-cue relationships are of comparable magnitude to the strongest correlations observed for English stops (Chodroff and Wilson, 2017, 2018; Sonderegger et al., 2020), demonstrating that structured speaker variability is present in laryngeal systems beyond English aspiration-type systems, and in more than one independent cue to a contrast in spontaneous speech.

Here, most of the predictable variability across individual speakers is *within* a given phonetic cue (IV A), as compared with variability *across* the two cues (IV B): no across-cue relationship (Table II) is as strong as either of the within-cue correlations (Table I). The size of the voicing contrasts between pVOT and VDC is weakly positively correlated across

speakers (Figure 4). This could be evidence that speakers vary in the degree of 'clarity' in their speech: speakers align multiple cues to a voicing contrast simultaneously in order to maximise the acoustic distinctiveness between the categories, as opposed to emphasising one cue over another (Bang, 2017; Clayards, 2018). An explanation in terms of speech clarity does not straightforwardly apply in this data, however, for two reasons. First, the size of the correlation itself is small (Table II, row 1), reflecting only a weak relationship between the two cue contrast sizes. Second, when comparing this to within-category relationships, this predictive pattern for the use of pVOT and VDC is observed only for voiced stops: whilst the pVOT-VDC relationship is negatively correlated in voiced stops, no clear relationship is observed for voiceless stops (Table II; Figure 5). This suggests that the pVOT-VDC cue relationship is *asymmetric* between stop voicing categories. This observation may indicate a restriction on structured speaker variability for only those segments in a series (i.e., voiced and voiceless stops) that have some form of featural specification. It has been previously argued that Japanese is a 'voiced' language (Ito and Mester, 1995; Mester and Ito, 1989; Nasukawa, 2005) in its being specified exclusively for a monovalent [voice] feature on voiced stops, with no featural specification for voiceless stops (e.g., Iverson and Salmons, 1995; Salmons, 2019). Furthermore, the lack of an observed correlation across cues may suggest that pVOT and VDC do not share an intrinsic link, potentially reflecting different articulatory pressures on their usage. This may be contrasted with stronger across-cue relationships between pVOT and closure voicing in Scottish English (Sonderegger et al., 2020). The lack of a correlation observed for Japanese, however, does not rule out a relationship between the cues: it is possible that VDC and pVOT, as measured here, simply do not capture the

dimensions in which these cues may be related. An alternative implementation of a closure voicing measure, distinct from the binary approach taken in this study, might reveal different across-cue patterns across different voicing specifications for stops. This empirical question would be an important direction for future research.

The within-cue findings (Section IV A) suggest that speakers can use cues independently to mark a linguistic contrast *without* maintaining the same cross-category relationships across more than one phonetic cue. This supports a restricted form of structured variability, constraining the predictability of speakers of spontaneous Japanese in their realisation of phonological categories along a single phonetic dimension. Crucially, speakers use two cues to *separately* realise the same phonological contrast. In this sense, the structured variability is *constrained*: here, speaker variability is present within a single acoustic cue, but speakers are less consistent in simultaneous use of multiple cues to the stop voicing contrast.

When considered from the perspective of a 'principle of uniformity' constraining phonetic variation (Chodroff and Wilson, 2017, 2018), our results provide some evidence for uniformity across speakers: namely, speakers are highly consistent *within* cues in signalling stop voicing contrasts. Our findings also demonstrate that a principle of uniformity is likely subject to constraints: here we find evidence of speakers covarying within individual cues, as opposed to covarying across more than one cue in marking the same contrast. Japanese differs from English in how the stop voicing contrast is specified: Japanese maintains a 'hybrid' stop voicing system involving the use of both positive VOT and voicing during closure (e.g., Nasukawa, 2005). Thus our evidence for covariation from Japanese stop voicing suggests that phonetic uniformity is constrained by language-specific properties. Our study emphasises

the importance of examining the evidence for uniformity in a range of empirical contexts, and especially across languages which differ in their phonetic implementation of a given phonological contrast.

A final point is that, despite distinct patterns in variability in pVOT and VDC observed across speakers, we did not observe age-graded differences in the use of these cues for marking the stop voicing contrast. Given a number of studies reporting a sound change towards an aspiration-based stop contrast (e.g., Gao et al., 2019; Takada, 2011; Takada et al., 2015), we may have expected to see an overall reduction in VDC in younger speakers. There are several reasons why we failed to observe this effect. The youngest speakers in this study were born during the 1970s, when this change was first observed, but the loss of prevoiced stops occurred later in the Tokyo region, where the speakers are from (Takada, 2011): our data may simply predate the widespread diffusion of the change. Alternatively, these differences may be obscured by using a binary implementation of VDC and controlling for the linguistic and social factors in the statistical model. Re-examining these questions with more recent data is an interesting direction for future research.

VI. CONCLUSION

This study has examined stops in spontaneous Japanese and demonstrated that structured variability is present in a new empirical setting, and that it is constrained in ways not straightforwardly predicted from studies mainly focussing on English. Specifically, the constraint arises from the linguistic specification and phonetic implementation of stop voicing in Japanese which requires a different configuration of acoustic cues from English. Such

497 a finding motivates an expanded search for structured speaker variability across more lan-
498 guages and phonetic cues. Within Japanese, for example, this could mean including F0 as
499 an acoustic cue, given its increasing importance for the stop voicing contrast (Gao and Arai,
500 2019; Gao et al., 2019; Kong et al., 2014). Our study provides the first sketch for a more
501 complex appreciation of how speaker variability is structured. It also motivates increasing
502 the range of studies on structured variability across languages, cues, and contrasts (Bang,
503 2017; Hauser, 2019; Hullebus et al., 2018).

504 ACKNOWLEDGMENTS

505 This paper is an extended version of a preliminary report in Tanner et al. (2019). We
506 thank two anonymous reviewers, Eleanor Chodroff, Meghan Clayards, Shigeto Kawahara,
507 James Kirby, and the audience of the 19th International Congress of Phonetic Sciences
508 (ICPhS) for their feedback on this research, Oriana Kilbourn-Ceron for programming as-
509 sistance, and Yuya Kawamata for translation assistance. Resources for statistical comput-
510 ing were provided by Calcul Québec and Compute Canada. This research reported was
511 supported by the Social Sciences and Humanities Research Council of Canada (#435-2017-
512 0925).

APPENDIX: POPULATION-LEVEL EFFECTS (PVOT)

Predictor	$\hat{\beta}$	Error	2.5% CrI	97.5% CrI
Intercept	3.11	0.02	3.08	3.15
Voicing	-0.51	0.02	-0.54	-0.48
Gender	-0.09	0.03	-0.15	-0.03
Previous phoneme manner (long)	0.03	0.00	0.03	0.04
Previous phoneme manner (nasal)	0.03	0.00	0.02	0.04
Birth year (1960-69)	0.04	0.02	-0.01	0.09
Birth year (1950-59)	0.03	0.02	-0.02	0.08
Birth year (1940-49)	0.00	0.03	-0.06	0.06
Birth year (1930-39)	-0.02	0.04	-0.09	0.05
Place of articulation (alveolar)	-0.18	0.01	-0.20	-0.15
Place of articulation (velar)	-0.12	0.01	-0.14	-0.10
Speech style (public speaking)	-0.10	0.00	-0.11	-0.09
Style style (dialogue)	0.01	0.00	0.00	0.02
Break Index (2)	0.05	0.00	0.05	0.06
Break Index (3)	0.05	0.00	0.04	0.05
Frequency (log)	-0.04	0.01	-0.05	-0.03
Speech rate (mean)	-0.06	0.03	-0.12	0.01
Speech rate (local)	-0.03	0.00	-0.04	-0.02
Preceding pause	0.04	0.01	0.02	0.05
Vowel height	0.14	0.01	0.11	0.16
Voicing : Gender	0.08	0.02	0.03	0.12
Voicing : Previous phoneme manner (long)	0.02	0.01	0.01	0.03
Voicing : Previous phoneme manner (nasal)	0.02	0.01	0.00	0.04
Voicing : Birth year (1960-69)	-0.03	0.02	-0.07	0.00
Voicing : Birth year (1950-59)	-0.05	0.02	-0.08	-0.01
Voicing : Birth year (1940-49)	0.04	0.02	0.00	0.08
Voicing : Birth year (1930-39)	-0.03	0.03	-0.08	0.02
Voicing : Place of articulation (alveolar)	0.05	0.02	0.02	0.09
Voicing : Place of articulation (velar)	0.06	0.01	0.04	0.09
Voicing : Speech style (public speaking)	0.03	0.01	0.01	0.04
Voicing : Speech style (dialogue)	-0.02	0.01	-0.03	0.00
Voicing : Break Index (2)	-0.06	0.01	-0.07	-0.05
Voicing : Break Index (3)	-0.04	0.00	-0.05	-0.03
Voicing : Frequency (log)	0.01	0.01	-0.01	0.04
Voicing : Speech rate (mean)	0.05	0.03	0.00	0.10
Voicing : Speech rate (local)	0.00	0.01	-0.02	0.01
Voicing : Preceding pause	-0.06	0.02	-0.10	-0.03
Voicing : Vowel height	-0.08	0.02	-0.13	-0.03

TABLE III. Estimate ($\hat{\beta}$), error, and 95% credible intervals for all population-level ('fixed effect') predictors for log-transformed pVOT.

APPENDIX: POPULATION-LEVEL EFFECTS (VDC)

Predictor	$\hat{\beta}$	Error	2.5% CrI	97.5% CrI
Intercept	-1.13	0.12	-1.36	-0.90
Voicing	2.99	0.14	2.72	3.25
Gender	0.12	0.18	-0.23	0.48
Previous phoneme manner (long)	0.01	0.03	-0.06	0.07
Previous phoneme manner (nasal)	-0.17	0.05	-0.27	-0.08
Birth year (1960-69)	0.33	0.14	0.04	0.61
Birth year (1950-59)	0.40	0.15	0.10	0.69
Birth year (1940-49)	-0.01	0.18	-0.35	0.34
Birth year (1930-39)	-0.36	0.21	-0.77	0.06
Place of articulation (alveolar)	0.00	0.07	-0.14	0.13
Place of articulation (velar)	0.13	0.05	0.04	0.22
Speech style (public speaking)	0.13	0.04	0.04	0.21
Speech style (dialogue)	-0.42	0.05	-0.52	-0.33
Break Index (2)	0.39	0.03	0.32	0.45
Break Index (3)	0.53	0.02	0.49	0.58
Frequency (log)	0.17	0.04	0.09	0.26
Speech rate (mean)	-0.57	0.19	-0.95	-0.20
Speech rate (local)	-0.16	0.04	-0.23	-0.09
Preceding pause	-3.24	0.16	-3.56	-2.93
Vowel height	0.12	0.07	-0.02	0.26
Voicing : Gender	0.06	0.20	-0.34	0.45
Voicing : Previous phoneme manner (long)	-0.20	0.06	-0.32	-0.07
Voicing : Previous phoneme manner (nasal)	-0.09	0.07	-0.22	0.04
Voicing : Birth year (1960-69)	-0.03	0.17	-0.36	0.29
Voicing : Birth year (1950-59)	0.04	0.17	-0.30	0.38
Voicing : Birth year (1940-49)	0.05	0.20	-0.34	0.44
Voicing : Birth year (1930-39)	-0.32	0.24	-0.78	0.14
Voicing : Place of articulation (alveolar)	0.24	0.12	0.01	0.46
Voicing : Place of articulation (velar)	0.13	0.09	-0.04	0.31
Voicing : Speech style (public speaking)	0.52	0.07	0.38	0.67
Voicing : Speech style (dialogue)	0.13	0.08	-0.03	0.28
Voicing : Break Index (2)	-0.54	0.06	-0.64	-0.42
Voicing : Break Index (3)	-0.57	0.03	-0.63	-0.50
Voicing : Frequency (log)	0.14	0.08	-0.02	0.30
Voicing : Speech rate (mean)	0.11	0.22	-0.31	0.55
Voicing : Speech rate (local)	0.10	0.06	-0.02	0.22
Voicing : Preceding pause	2.00	0.21	1.58	2.41
Voicing : Vowel height	0.60	0.15	0.31	0.90

TABLE IV. Estimate ($\hat{\beta}$), error, and 95% credible intervals for all population-level ('fixed effect') predictors for VDC (logit-scale).

[1]pVOT was log-transformed in order to meet the the assumption of linear regression that the response is a linear function of the parameters, and to account for non-normality in the distribution.

[2]To ensure that the correlations reported were not due to the choice of a specific prior, an identical model with a weaker 'flat' prior ($\zeta = 1$) was also fit. The correlations estimated from this model, of primary interest for our research questions, were near identical (within 0.01) to those from the stronger model, indicating that the evidence for the correlations in the data is strong enough not to be affected by the subjective choice to use a more informative prior.

Abramson, A. S. and Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *J. Phon.*, 63:75–86.

Allen, S. J., Miller, J. L., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *J. Acoust. Soc. Am.*, 113:544–552.

Bang, H.-Y. (2017). *The structure of multiple cues to stop categorization and its implications for sound change*. PhD thesis, McGill University.

Baran, J., Laufer, M., and Daniloff, R. (1977). Phonological contrastivity in conversation: a comparative study of voice onset time. *J. Phon.*, 5:339–350.

Beckman, J., Jessen, M., and Ringen, C. (2013). Empirical evidence for laryngeal features: aspirating vs. true voicing languages. *Journal of Linguistics*, 49:259–284.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer (version 6.0.36).

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Bybee, J. B. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Cho, T. and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *J. Phon.*, 27:207–229.

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *J. Phon.*, 61:30–47.

Chodroff, E. and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, 4.

Clayards, M. (2018). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, 75:1–23.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, Hillsdale, NJ.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *J. Phon.*, 54:35–60.

Davidson, L. (2018). Phonation and laryngeal specification in American English voiceless obstruents. *Journal of the International Phonetic Association*, 48:331–356.

DiCanio, C., Nam, H., Amith, J. A., Garcia, R. C., and Whalen, D. H. (2015). Vowel variablility in elicited versus spontaneous speech: evidence from Mixtec. *J. Phon.*, 48:45–59.

Docherty, G. (1992). *The timing of voicing in British English obstruents*. Foris, Berlin & New York.

Eager, C. (2015). Automated voicing analysis in Praat: statistically equivalent to manual segmentation. In *Proc. 18th ICPhS* . University of Glasgow.

Foulkes, P., Docherty, G., and Watt, D. (2001). The emergence of structured variation. *Univ. Penn. Work. Papers. Ling.*, 7:67–84.

Fujimoto, M., Kikuchi, H., and Maekawa, K. (2006). Corpus of Spontaneous Japanese documentation: phone information. Technical Report 6, National Institute for Japanese Language and Linguistics, Tokyo.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66:789–806.

Gao, J. and Arai, T. (2019). Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: positional variation, cue enhancement, and contrast recovery. *J. Phon.*, 77:1–33.

Gao, J., Yun, J., and Arai, T. (2019). VOT and F0 coarticulation in Japanese: production-biased or misparsing? In *Proc. 19th ICPhS* , Melbourne, Australia.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge.

Hauser, I. (2019). *Effects of Phonological Contrast on Within-Category Phonetic Variation*. PhD thesis, University of Massachusetts Amherst.

Hullebus, M. A., Tobin, S. J., and Gafos, A. I. (2018). Speaker-specific structure in German voiceless stop voice onset times. In *Proc. Interspeech*, pages 1403–1407, Hyderabad.

Hunnicutt, L. and Morris, P. A. (2016). Prevoicing and aspiration in Southern American English. In *Proceedings of the 39th Annual Penn Linguistics Conference*, volume 22, Philadelphia. University of Pennsylvania.

Ito, J. and Mester, A. R. (1995). Japanese phonology. In Goldsmith, J. A., editor, *The Handbook of Phonological Theory*, pages 817–838. Blackwell.

Iverson, G. and Salmons, J. (1995). Aspiration and laryngeal representation in Germanic. *Phonology*, 12:369–396.

Johnson, K., Ladefoged, P., and Lindau, M. (1993). Individual differences in vowel production. *J. Acoust. Soc. Am.*, 94:701–714.

Kawahara, S. (2015). Geminate devoicing in Japanese loanwords: Theoretical and experimental investigations. *Language and Linguistics Compass*, 9:181–195.

Kay, M. (2019). *tidybayes: Tidy Data and Geoms for Bayesian Models*. R package version 1.0.4.

Kikuchi, H. and Maekawa, K. (2003). Performance of segmental and prosodic labeling of spontaneous speech. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo. Tokyo Institute of Technology.

Kim, S., Kim, J., and Cho, T. (2018). Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *J. Phon.*, 71:65–80.

Klatt, D. (1975). Voice onset time, frication and aspiration in word-initial consonant clusters. *J. Speech Lang. Hear. Res.*, 18:686–706.

Kleber, F. (2018). VOT or quantity: What matters more for the voicing contrast in German regional varieties? results from apparent-time analyses. *J. Phon.*, 71:468–486.

Kleinschmidt, D. F. (2018). Structure in talker variability: How much is there and how much can it help? *Language, Cognition, and Neuroscience*, 34:1–26.

Kong, E. J., Yoneyama, K., and Beckman, M. E. (2014). Effects of a sound change in progress on gender-marking cues in Japanese. In *Proceedings of LabPhon 14*, Tokyo. NINJAL.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100:1989–2001.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1:153–167.

Liberman, M. A., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74:431–461.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, volume 4 of *NATO ASI Series*, pages 403–439. Kluwer Academic Publishers.

Lisker, L. (1986). Voicing in English: a catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech*, 29:3–11.

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422.

Lisker, L. and Abramson, A. S. (1967). Some effects of context on voice onset time in English. *Language and Speech*, 10:1–28.

623 Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. (2002). X-JToBI: An extended
624 J_ToBI for spontaneous speech. In *Proc. 7th Int. Conf. Spoken Language Processing*, pages
625 1545–1548, Denver.

626 Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of
627 Japanese. In *Proc. 2nd LREC*, volume 2, pages 946–952.

628 Mester, A. and Ito, J. (1989). Feature predictability and underspecification: palatal prosody
629 in Japanese mimetics. *Language*, 65:258–293.

630 Meunier, C. and Espresser, R. (2011). Vowel reduction in casual French: the role of lexical
631 factors. *J. Phon.*, 39:271–278.

632 Nasukawa, K. (2005). The representation of laryngeal-source contrasts in Japanese. In van de
633 Weijer, J., Nanjo, K., and Nishihara, T. ., editors, *Voicing in Japanese*, pages 71–87. De
634 Gruyter Mouton.

635 Nicenboim, B. and Vasishth, S. (2016). Statistical methods for linguistic research: Founda-
636 tional ideas - part II. *Language and Linguistics Compass*, 10:591–613.

637 Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels.
638 *J. Acoust. Soc. Am.*, 24:175–184.

639 Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lentition, and contrast.
640 In Bybee, J. and Hopper, P., editors, *Frequency and the Emergence of Linguistic Structure*,
641 pages 137–157. John Benjamins.

642 Riney, T. J., Takagi, N., Ota, K., and Uchida, Y. (2007). The intermediate degree of VOT
643 in Japanese initial stops. *J. Phon.*, 35:439–443.

644 Salmons, J. (2019). Laryngeal phonetics, phonology, assimilation and final neutralization.
645 In Page, R. and Putnam, M. T., editors, *Cambridge Handbook of Germanic Linguistics*,
646 pages 119–142. Cambridge University Press, Cambridge.
647 Schertz, J., Cho, T., Lotto, A., and Warner, N. (2015). Individual differences in phonetic
648 cue use in production and perception of a non-native sound contrast. *J. Phon.*, 52:183–204.
649 Schultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception
650 and production of consonant voicing. *J. Acoust. Soc. Am.*, 132.
651 Seyfarth, S. and Garellek, M. (2018). Plosive voicing acoustics and voice quality in Yerevan
652 Armenian. *J. Phon.*, 71:425–450.
653 Shimizu, K. (1996). *A cross-language study of the voicing contrasts of stop consonants in*
654 *Asian languages*. Seibido, Tokyo.
655 Sonderegger, M., Bane, M., and Graff, P. (2017). The medium-term dynamics of accents on
656 reality television. *Language*, 93:598–640.
657 Sonderegger, M., Stuart-Smith, J., Knowles, T., MacDonald, R., and Rathcke, T. (2020).
658 Structured heterogeneity in Scottish stops over the twentieth century. *Language*, 96:94–125.
659 Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R. (2015). The private life
660 of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Lab. Phon.*, 6:505–549.
661 Takada, M. (2011). *Nihongo no gotou heisa'on no kenkyuu: VOT no kyoujiteki bunpu to*
662 *tsuujiteki henka [Research on the word-initial stops of Japanese: synchronic distribution*
663 *and diachronic change in VOT]*. Kurosio, Tokyo.
664 Takada, M., Kong, E. J., Yoneyama, K., and Beckman, M. E. (2015). Loss of prevoicing in
665 Modern Japanese /g, d, b/. In *Proc. 18th ICPhS* . University of Glasgow.

666 Tanner, J., Sonderegger, M., and Stuart-Smith, J. (2019). Structured speaker variability
667 in spontaneous Japanese stop contrast production. In *Proc. 19th ICPhS* , Melbourne,
668 Australia.

669 Theodore, R. M., Miller, J. L., and DeSteno, D. (2009). Individual talker differences in
670 voice-onset-time: Contextual influences. *J. Acoust. Soc. Am.*, 126:3974–3982.

671 Tsujimura, N. (2014). *Introduction to Japanese Linguistics*. Wiley-Blackwell, Oxford.

672 Vasishth, S., Nicenboim, B., Beckman, M., Li, F., and Kong, E. J. (2018). Bayesian data
673 analysis in the phonetic sciences: A tutorial introduction. *J. Phon.*, 71:147–161.

674 Venditti, J. (2005). The J_ToBI model of Japanese intonation. In Sun-Ah, J., editor,
675 *Prosodic Typology*, pages 172–200. Oxford University Press, Oxford.

676 Yao, Y. (2009). Understanding VOT variation in spontaneous speech. *UC Berkeley Phonol-*
677 *ogy Lab Annual Report*, pages 29–43.